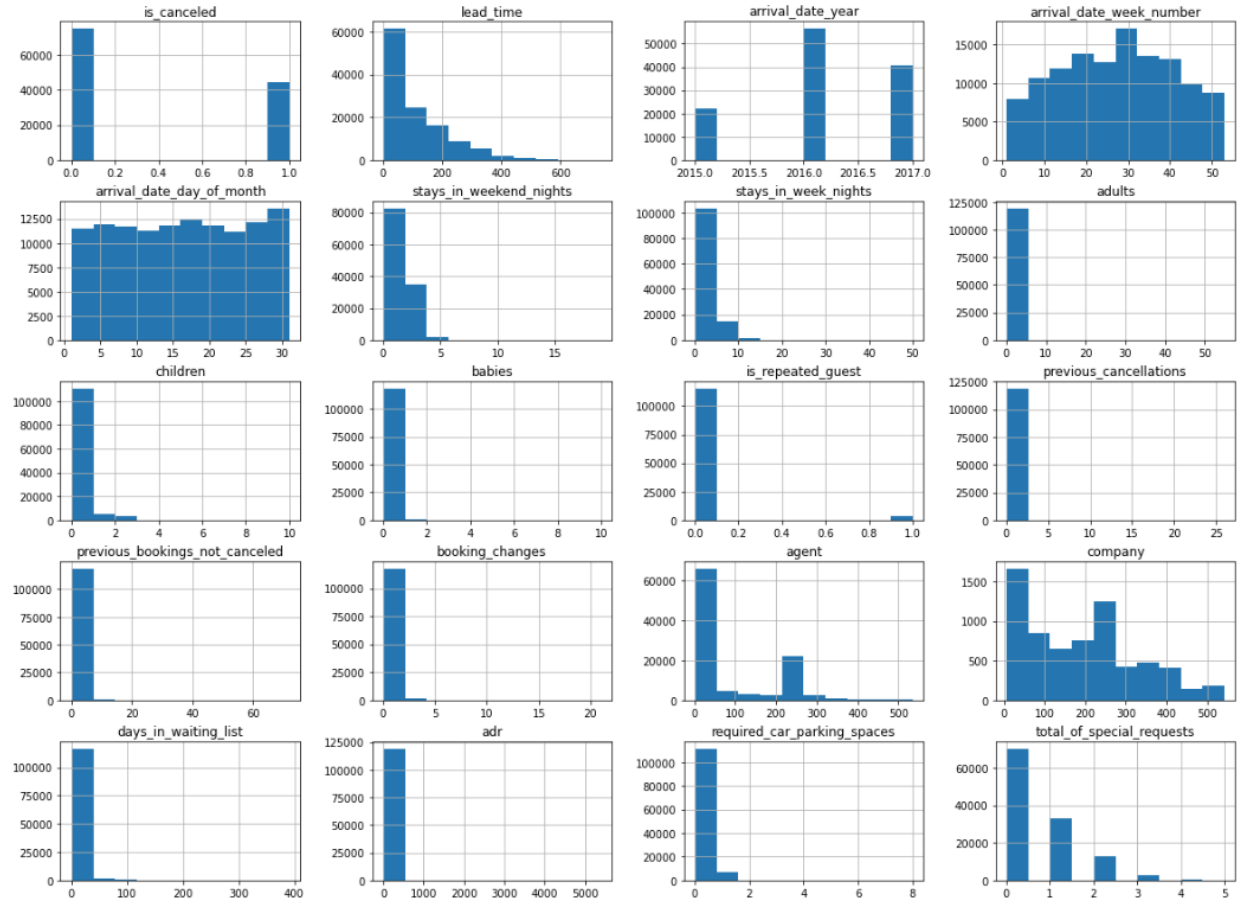
Hotel Booking Analysis and Forecasting

1. Project Background

The global pandemic has impacted all industries, with the hospitality sector experiencing some of the most significant challenges. As the world transitions into a post-pandemic phase, changes both minor and major are emerging within this sector. Signs of this transformation are visible, including governments authorizing the resumption of both domestic and international air travel. This development encourages individuals who have remained isolated for a considerable period to consider once again traveling for leisure. Consequently, analyzing the hotel reservation data, forming a cognition for the hotel business, finding key indicators, and making suggestions on how to further promote the hotel reservation service and build a hotel potential customer identification model is critical.

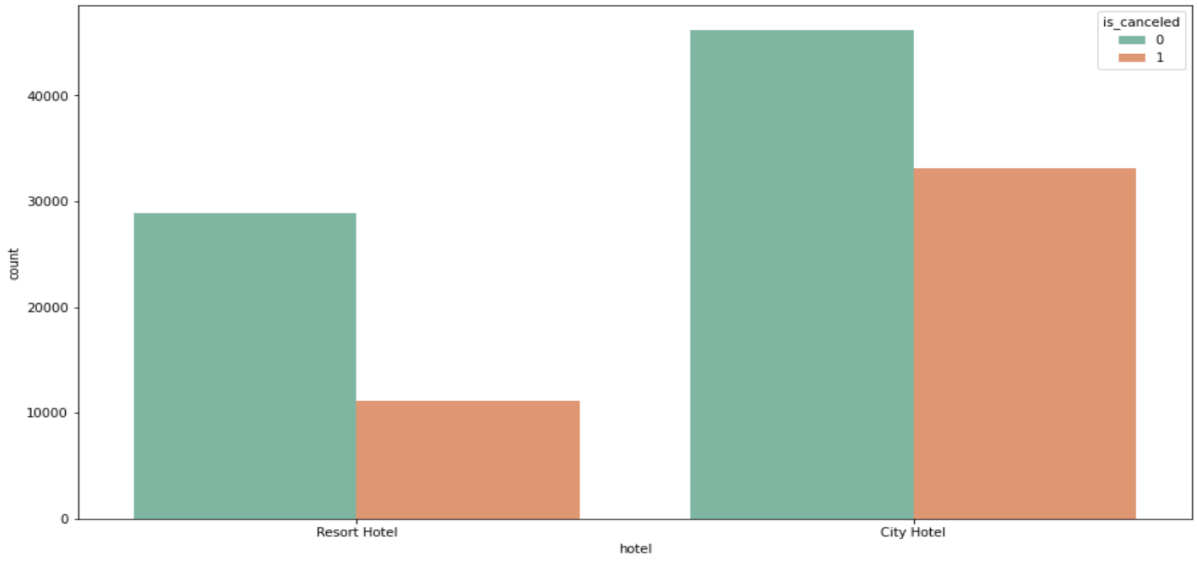
2. Data Collection and Preparation

This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

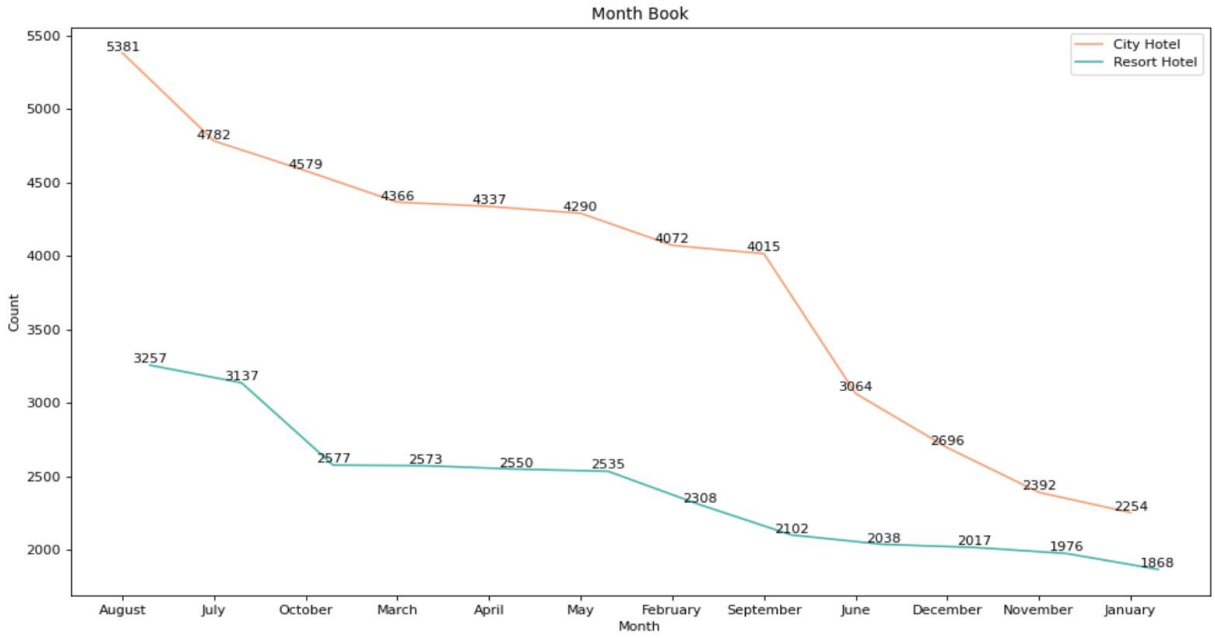


3. Main Exploratory Data Analysis (EDA):

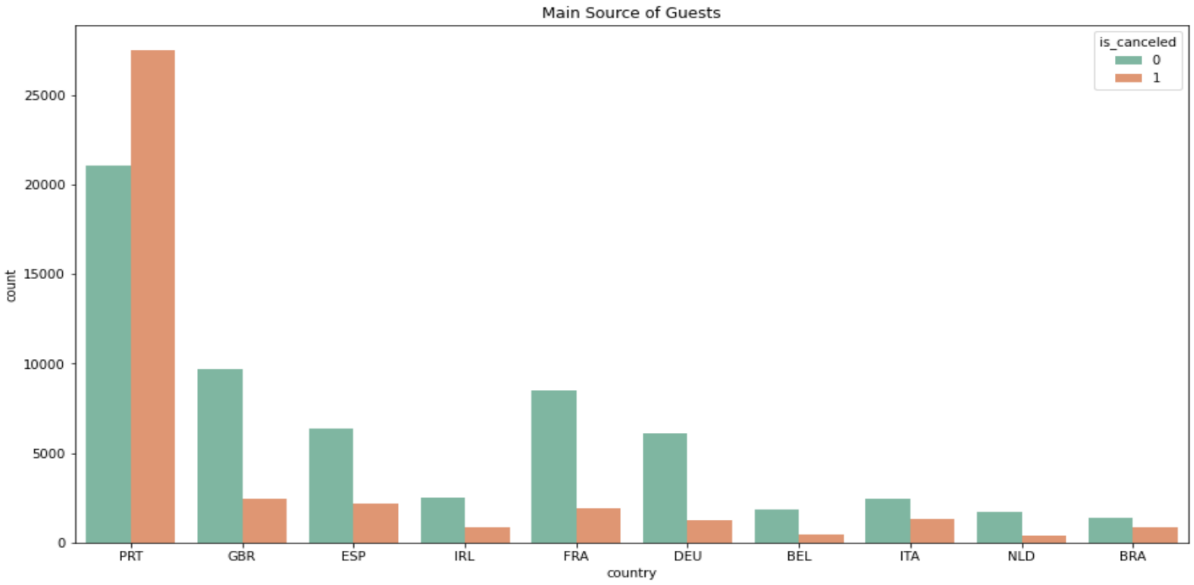
1. Hotel bookings and cancellations: City Hotel's booking volume and cancellation volume are both higher than Resort Hotel's, but Resort Hotel's cancellation rate is 27.8%, while City Hotel's cancellation rate reaches 41.7%.



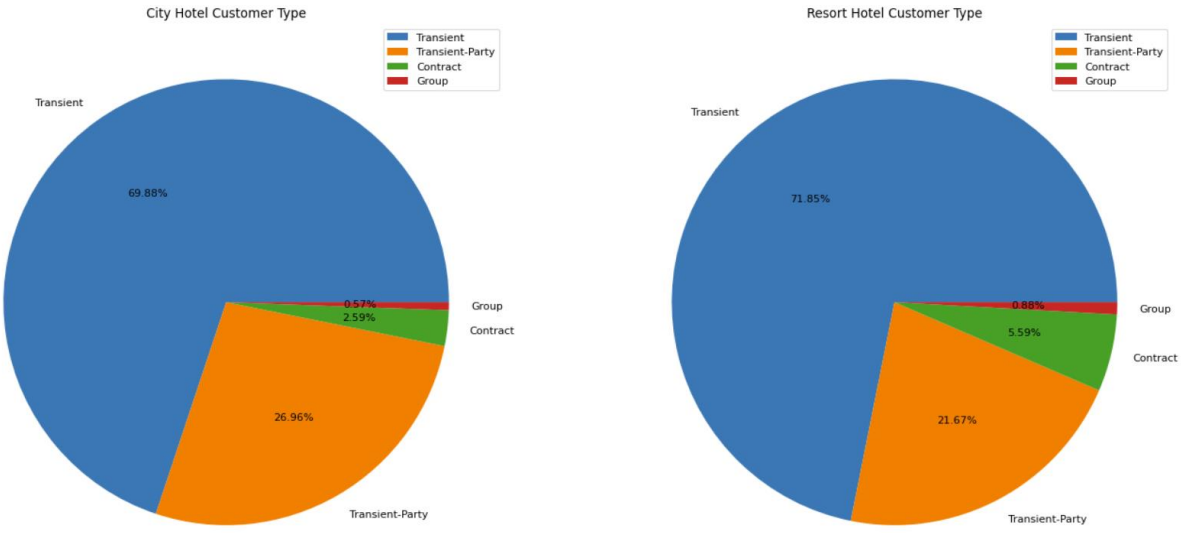
2. Hotel bookings by month: Peak booking months are August and July. Preliminary judgment is that the long holiday caused the peak period.



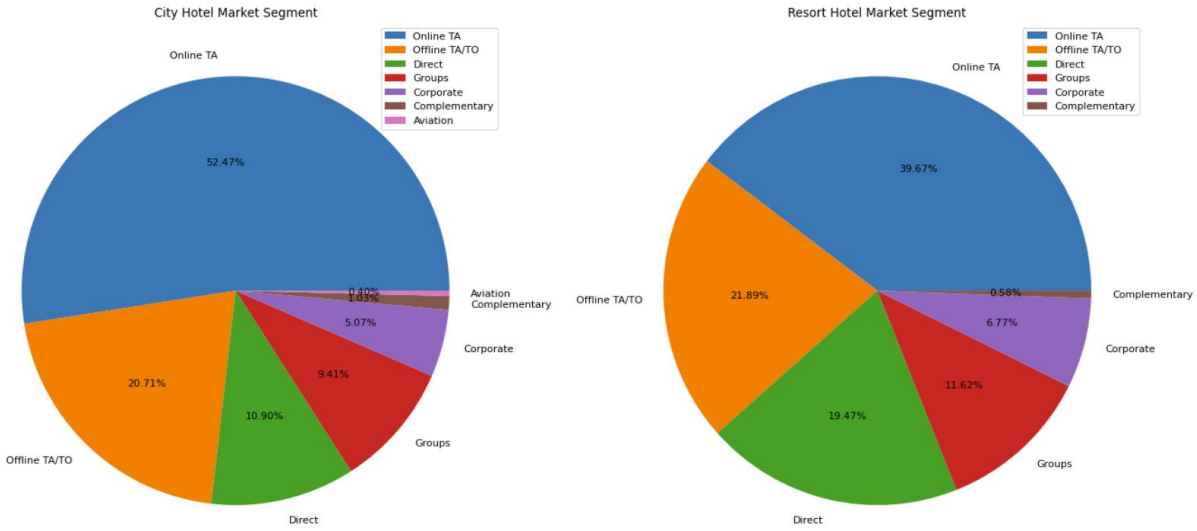
3. Customer origin and booking cancellation rate: The peak season for both Resort hotel and City hotel is July and August in summer, and the main sources of tourists are European countries. This is in line with the characteristics of European tourists who prefer summer travel. It is necessary to focus on countries with high cancellation rates such as Portugal (PRT) and the United Kingdom (BRT). Main source of customers.



4. Customer Type: The main customer type of the hotel is transient travelers, accounting for about 70%.



5. Hotel Booking Method: The customers of the two hotels mainly come from online travel agencies, which account for even more than 50% of the City Hotel; offline travel agencies come next, accounting for about 20%.



4. Feature Engineering and Selection:

a. Handling Categorical Variables: The project identified categorical features within the data, including those expressed numerically like 'agent', 'company', and 'is_repeated_guest'. These features were appropriately encoded for model readiness.

b. Creating New Features: New variables like 'in_company', 'in_agent', and 'same_assignment' were introduced. These features were designed to capture additional nuances in the data - for instance, 'same_assignment' indicated whether the booked room type matched the assigned room type.

c. Feature Removal: Certain features, such as 'reserved_room_type', 'assigned_room_type', 'agent', and 'company', were dropped to streamline the model and focus on more impactful variables.

d. Working with Continuous Variables: Continuous variables were identified and standardized, ensuring they were appropriately scaled for model training.

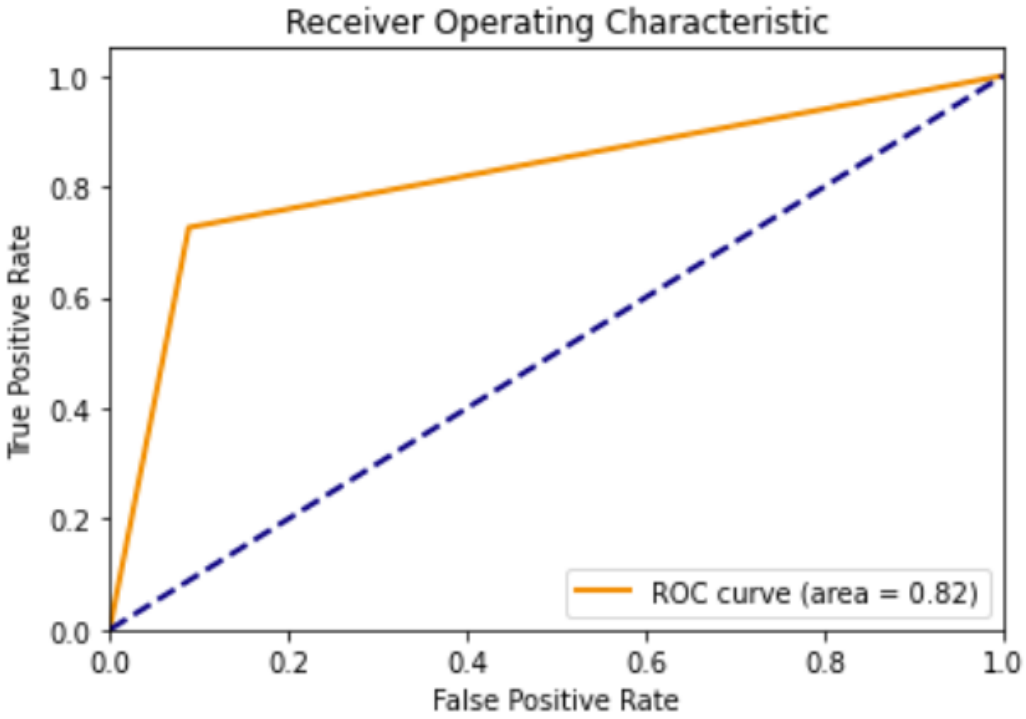
e. Correlation Analysis: The correlation of numerical columns with the target variable 'is_canceled' was calculated to understand the relationship and influence of each feature on the prediction outcome.

5. Model Building:

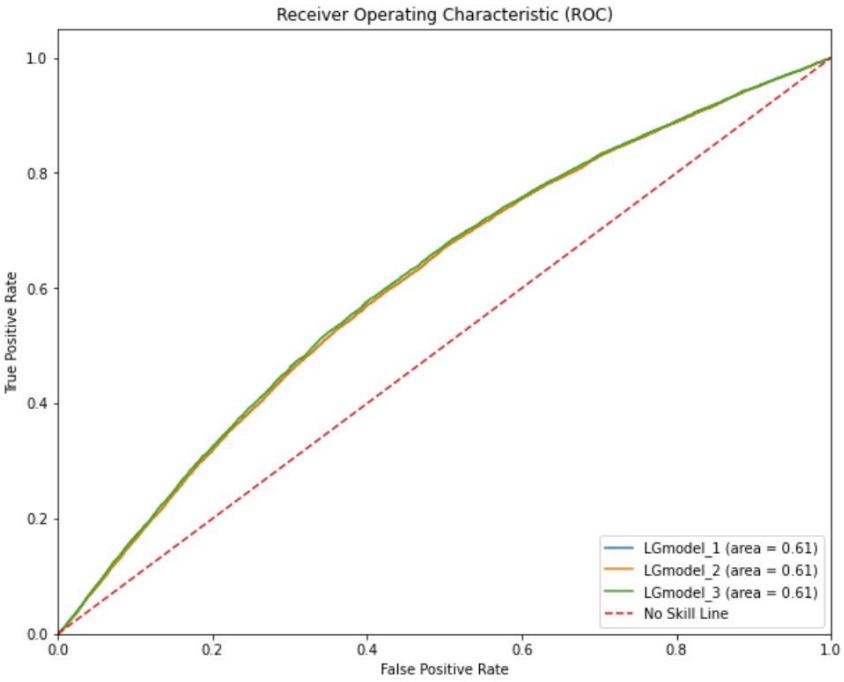
Key models explored include RandomForestClassifier, Logistic Regression, and XGBoost, each with different configurations and hyperparameters. The analysis includes evaluating the performance of these models using metrics like accuracy, precision, recall, and ROC curves. It also discusses the rationale

behind choosing specific models and configurations, based on their performance and suitability for the dataset's characteristics.

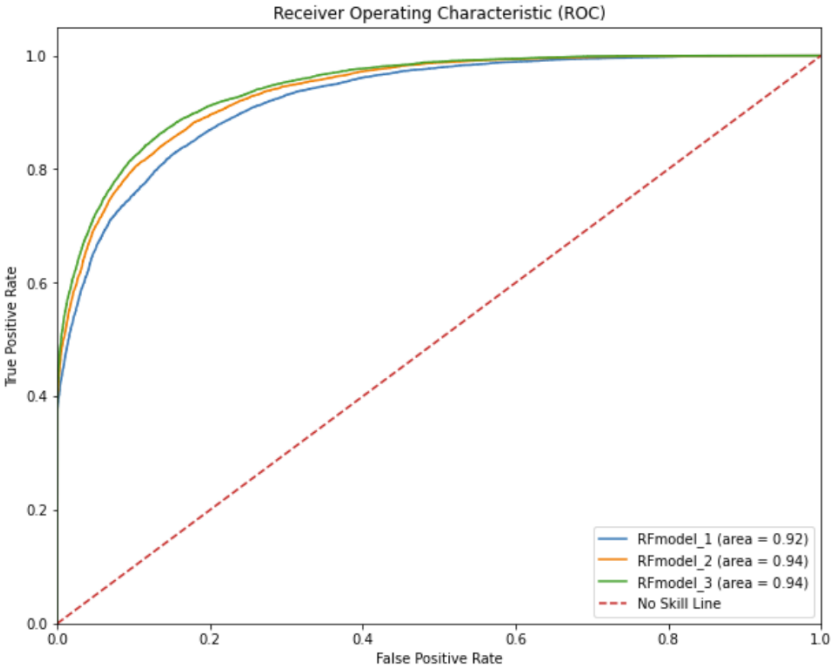
a. XGBoost: XGModel_1 shows a strong performance on the training set with an accuracy of 86.55% and an F1 score of 81.11%. Its ROC-AUC of 84.74% on the training set and 84.11% on the test set indicates good discriminative ability. XGModel_2 has a slightly lower training accuracy of 81.31% and an F1 score of 70.25%. The ROC-AUC scores are lower compared to XGModel_1, with a test ROC-AUC of 76.85%, suggesting that this model is less capable of distinguishing between the classes. XGModel_3 presents an improvement over XGModel_2 with a training accuracy of 84.41% and an F1 score of 77.61%. The ROC-AUC score is 82.01% for the training set and 81.87% for the test set, indicating a better balance between sensitivity and specificity than XGModel_2. The ROC curve provided for one of the models has an area under the curve (AUC) of 0.82, which is a strong score, suggesting the model has good classification abilities.



b. Logistic: The Logistic Regression models (LGmodel_1, LGmodel_2, LGmodel_3) appear to perform modestly with similar accuracy across all three versions, hovering around the 62-63% mark for both training and test sets. The precision metrics are also consistent, suggesting that the model's ability to predict true positives and overall positive predictions is stable. However, the recall is relatively low, indicating that the models miss a significant number of actual positives. The consistent ROC-AUC score of 0.61 across the three models indicates limited discrimination capacity between the positive and negative classes, which is just slightly better than random guessing (represented by the no-skill line with an AUC of 0.5).



c. RandomForest: The RandomForest models (RFmodel_1, RFmodel_2, RFmodel_3) demonstrate a progressive improvement in both training and test metrics. The training accuracy starts from 84.86% in RFmodel_1 and reaches 92.23% in RFmodel_3, showing enhanced learning as the model's progresses. Precision and recall metrics on test data indicate a growing ability to correctly predict cancellations without many false positives or negatives. The ROC curves reflect excellent predictive performance, with AUC scores well above the no-skill line, indicating strong discriminative ability. RFmodel_3 shows the highest AUC of 0.94, pointing to a superior balance between true positive rate and false positive rate.

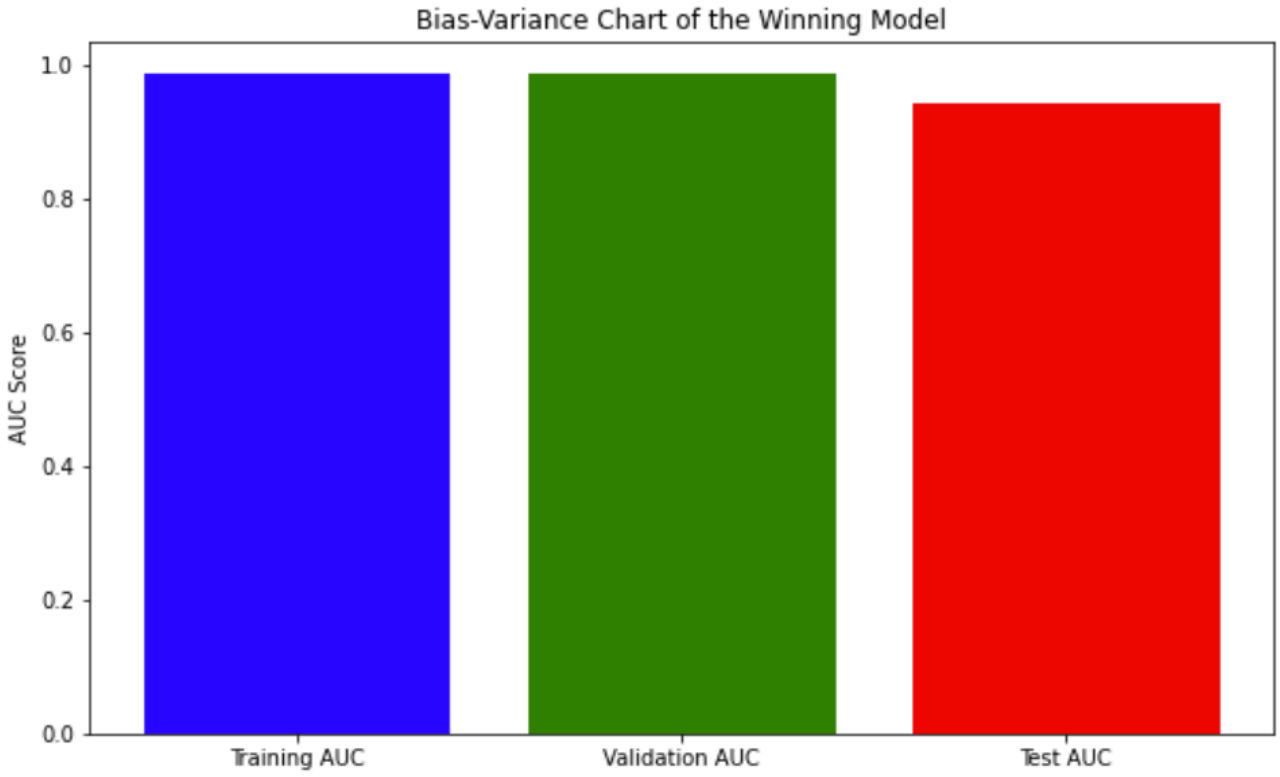


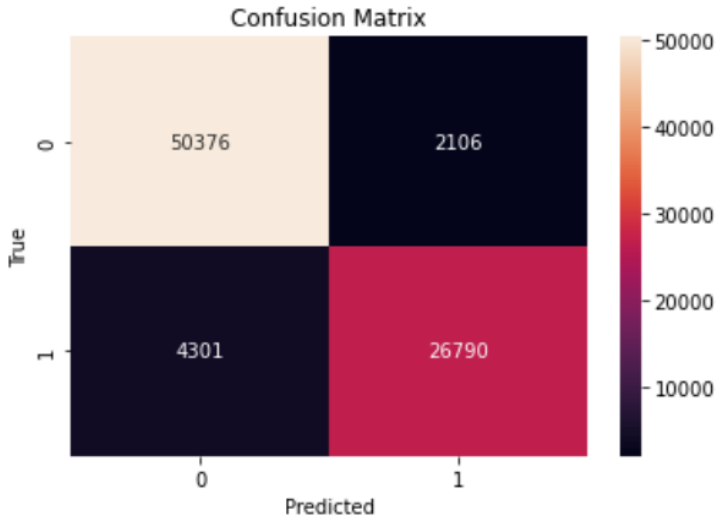
6. Model Evaluation:

The winning model, RFmodel_3, demonstrated a high validation AUC of approximately 0.986 and maintained strong performance with a test AUC of around 0.944. This high level of performance on both validation and test sets suggests that RFmodel_3 has a robust predictive capability and generalizes well to unseen data, making it a reliable choice for forecasting hotel booking cancellations. The code snippet shown indicates the process for obtaining these AUC scores and reporting the winning model's performance, showcasing the model's effectiveness in classification tasks relevant to the project's objectives.

7. Bias Detection:

The confusion matrix shows that the model has a significant number of both false positives and false negatives, indicating potential prediction bias towards certain classes. The ROC curve demonstrates that the model has reasonable discriminative ability. However, an AUC of 0.92-0.94 suggests that there's still room for improvement. The distribution of predicted probabilities reveals that the model may be overconfident with some predictions (probabilities close to 0 or 1), which could be indicative of bias in the model.





8. Mitigation Strategies:

From the feature importance chart, the 'total_stays' and 'same_assignment' are among the most influential features for predicting cancellations. This suggests that the length of the stay and whether the assigned room type matches the reserved room type are significant predictors of cancellation. The prediction probability charts show the model's confidence in its predictions, with the model sometimes being quite certain about a booking not being canceled and a 0.90 probability of 'Not Canceled' and other times less certain. The bar chart reflects accuracy metrics before and after bias mitigation strategies were applied. It shows similar levels of accuracy, indicating that the mitigation strategies did not adversely affect the model's ability to predict correctly.